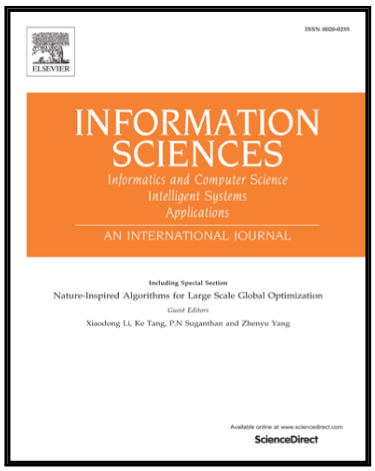


Exploiting Reject Option in Classification for Social Discrimination Control

Faisal Kamiran, Sameen Mansha, Asim Karim, Xiangliang Zhang

PII: S0020-0255(17)30983-0
DOI: 10.1016/j.ins.2017.09.064
Reference: INS 13168

To appear in: *Information Sciences*

Received date: 30 March 2016
Revised date: 1 September 2017
Accepted date: 27 September 2017

Please cite this article as: Faisal Kamiran, Sameen Mansha, Asim Karim, Xiangliang Zhang, Exploiting Reject Option in Classification for Social Discrimination Control, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.09.064

Exploiting Reject Option in Classification for Social Discrimination Control

Faisal Kamiran[a,*], Sameen Mansha[b,a], Asim Karim[c,a], Xiangliang Zhang[d]

[a] *Information Technology University, Lahore, Pakistan*
[b] *School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia*
[c] *Department of Computer Science, Syed Babar Ali School of Science and Engineering, Lahore University of Management Sciences, Lahore, Pakistan*
[d] *Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, KSA*

Abstract

Social discrimination is said to occur when an unfavorable decision for an individual is influenced by her membership to certain protected groups such as females and minority ethnic groups. Such discriminatory decisions often exist in historical data. Despite recent works in discrimination-aware data mining, there remains the need for robust, yet easily usable, methods for discrimination control. In this paper, we utilize reject option in classification, a general decision theoretic framework for handling instances whose labels are uncertain, for modeling and controlling discriminatory decisions. Specifically, this framework permits a formal treatment of the intuition that instances close to the decision boundary are more likely to be discriminated in a dataset. Based on this framework, we present three different solutions for discrimination-aware classification. The first solution invokes probabilistic rejection in single or multiple probabilistic classifiers while the second solution relies upon ensemble rejection in classifier ensembles. The third solution integrates one of the first two solutions with situation testing which is a procedure commonly used in the court of law. All solutions are easy to use and provide strong justifications for the

*Corresponding author
Email addresses: faisal.kamiran@itu.edu.pk (Faisal Kamiran), s.mansha@uqconnect.edu.au (Sameen Mansha), akarim@lums.edu.pk (Asim Karim), xiangliang.zhang@kaust.edu.sa (Xiangliang Zhang)

decisions. We evaluate our solutions extensively on four real-world datasets and compare their performances with previously proposed discrimination-aware classifiers. The results demonstrate the superiority of our solutions in terms of both performance and flexibility of applicability. In particular, our solutions are effective at removing illegal discrimination from the predictions.

Keywords: Discrimination-aware Data Mining; Fairness in Machine Learning; Classification; Decision Theory

1. Introduction

Social discrimination is said to occur when a decision in favor of or against a person is made based on the group, class, or category to which that person belongs to rather than on merit. Discriminatory practices suppress opportunities for members of deprived groups in employment, income, education, finance, and other benefits/services on the basis of their age, gender, skin color, religion, race, language, culture, marital status, economic condition, and other non-merit factors. Today, discrimination is considered unacceptable from social, ethical, and legal perspectives. Many anti-discrimination laws [3, 11, 28, 27] have been enacted and many anti-discrimination organizations (e.g., ENAR [1]) are working for the eradication of discrimination. The consequences of discriminatory practices can range from legal prosecution to a variety of social problems like high unemployment rate, frustration, low productivity, and social unrest.

The discrimination-aware classification problem studies the construction and application of classifiers learned from discriminatory or biased data. The do-nothing approach of simply using a classifier learned from discriminatory data will propagate, if not exacerbate, discriminatory decisions, which is undesirable for decision makers at financial institutions, hiring agencies, and social service providers. Thus, this do-nothing approach can lead to litigations and penalties.

In recent years, several methods have been proposed for discrimination-aware classification. However, these methods have one or both of the following shortcomings. First, they require that either the discriminatory data is processed to

remove discriminatory patterns before learning a classifier or a specific classifier's learning algorithm is modified to make it discrimination-aware. Second, they are usually 'brute force' techniques with limited control over overall and illegitimate (unexplainable) discrimination removal.

These shortcomings of existing methods have hindered their adoption by practitioners. A direct consequence of the first shortcoming is that whenever discrimination w.r.t. a different sensitive attribute needs to be addressed, the historical data or classifier needs to be processed again. Our experience with the *Dutch Research and Documentation Center* (WODC) associated with the Ministry of Security and Justice and *Statistics Netherlands*, the national census body, confirms the importance of tackling discrimination w.r.t. multiple factors including age, gender, and race [18]. Being restricted to a specific discrimination-aware classifier (e.g., naive Bayes [6], decision tree [17]) is also an issue because that classifier may not be the best performing classifier for a given dataset. The second shortcoming can lead to reverse discrimination whereby deprived group individuals are favored without a legitimate or plausible explanation. This issue has been studied by the authors of [32]. They split overall discrimination into legal and illegal parts and claim that if the discrimination (e.g., high income of male employees as compared to female employees) can be explained by some reasonable factors (e.g., longer working hours of males), then it is acceptable and legitimate 'discrimination' rather than illegal discrimination. On the other hand, it would be illegal to discriminate on the basis of sensitive factors (e.g., gender, race) without any plausible explanation. The current state-of-the-art methods either deal with the overall discrimination or illegal discrimination and are not flexible enough to prevent both overall and illegal discrimination simultaneously.

In this paper, we develop and evaluate a methodology for making single and ensembles of classifiers discrimination-aware w.r.t. overall and illegal discrimination. This methodology is based on the decision theoretic notion of reject option where instances with highly uncertain labels are not given one in classification (i.e., they are given the reject label). Previously, it has been

hypothesized that discriminatory decisions are often made close to the decision boundary because of decision maker's bias [16]. Our proposed methodology formalizes this into practically usable solutions for discrimination-aware classification. Furthermore, the rejected instances represent potentially discriminated or favored instances in the biased dataset. Thus, our methodology also serves as a model-based discrimination discoverer in biased datasets.

We present three rejection strategies and corresponding rules for discrimination control in predictions. The first solution called Probabilistic Rejection (PR), rejects instances with uncertain posterior probabilities, thus enabling it to be used with any probabilistic classifier or ensemble of classifiers. Our second rejection strategy, called Ensemble Rejection (ER), identifies instances that are not unanimously labeled by an ensemble of classifiers, thus emulating the natural decision making process by a group of experts. Our third rejection strategy, called Situational Rejection (SR), combines probabilistic rejection or ensemble rejection with situation testing to identify discriminated instances. Situation testing is a legally admissible procedure for verifying discrimination cases by comparing them with other similar cases. All strategies/solutions include relabeling rules with parametric control over the resulting discrimination. We perform extensive experiments to verify the superior performance of our methodology. In particular, we also demonstrate that our methodology prefers removing illegal discrimination over explainable discrimination while reducing overall discrimination. Thus, it addresses a common criticism that discrimination prevention methods disregard explainable discrimination while removing overall discrimination.

The rest of the paper is organized as follows. Section 2 discusses the related work in discrimination-aware classification. Section 3 defines the problem setting and measures for overall and illegal discrimination. We present our reject option based methodology and specific solutions in Section 4. Section 5 presents experimental evaluations and discussions of our solutions. We summarize and conclude our contribution in Section 6.

2. Related Work

Data mining techniques can assist with the discovery of discriminatory patterns from data and with preventing discriminatory decisions based on biased data. The topic of social discrimination in data mining was introduced by Pedreschi et al. in 2008 [24].Since then many researchers have focused on discrimination detection and prevention in data mining. A multidisciplinary survey of discrimination analysis methods is given by [25] while an edited book provides a summary of the research works for discrimination discovery and prevention [8]. The book also deals with the legal and ethical issues of discrimination and profiling.

Proposed methods for discrimination prevention requiring learning model adaptation include those for decision trees [17], naive Bayes classifiers [6], logistic regression [20], and support vector machines (SVM) [31]. All these methods require that the learning model or algorithm is tweaked, and these methods are specific to their respective classifiers. For example, in [17], the authors propose a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free while in [31] fairness constraints are introduced to control discrimination in discriminative classifiers like SVM.

Direct discrimination arises when sensitive attributes are utilized in learning and prediction. Nonetheless, it has been shown that discrimination is not removed by simply removing these attributes from the dataset [16]. That is, discriminatory decisions can still be made due to correlation of sensitive attributes with other attributes (indirect discrimination or *redlining*[1]. This issue has been studied in greater detail in [32]. The authors of [32] also present the concept of explainable and illegal discrimination and propose a variant of data preprocessing approaches of [16] to prevent the illegal discrimination only. However, their method is unable to handle multiple explanatory attributes and both explainable and illegal discrimination simultaneously. More recently, propensity score

[1] `http://en.wikipedia.org/wiki/Redlining`, March. 12, 2016

modeling has been introduced by [7] to filter out illegal discrimination from data. Subsequently, they develop analytical solutions for discrimination-aware linear regression that controls the illegal effect of an attribute on the outcome.

In our previous work [19], we presented two strategies for making standard classifiers and classifier ensembles discrimination-aware at run-time. Based on decision theory, these strategies provided stronger control and interpretability of the decisions. A similar approach of shifting the decision boundary has been shown by [12] to produce good accuracy-discrimination trade-off performance. In this paper, we generalize our strategies to a model of discrimination based on reject option in classification. This model leads to a methodology for discrimination control in predictions. Following this methodology, we present three solutions for discrimination control, including a new solution incorporating situation testing, and evaluate them extensively for both illegal and overall discrimination prevention. These solutions require neither data preprocessing nor algorithm tweaking, and can be utilized with a variety of classifiers with ease.

3. Background and Notation

This section defines the problem setting and introduces the measures used in this work.

3.1. Problem Definition

We consider a two-class classification problem with label $C \in \{C^+, C^-\}$ defined over instances $X \in \mathcal{X}$ described by a fixed number of attributes. A discriminatory dataset $\mathcal{D} = \{X_i, C_i\}_{i=1}^N$ is available in which the labels C_i are biased w.r.t. one or more sensitive or discriminatory attributes S, e.g., Gender or Race. We assume that C^+ is the desirable label. The instances in $\mathcal{X}$ can be distinguished between those belonging to a deprived group $\mathcal{X}^d$ or a favored group $\mathcal{X}^f$, where $\mathcal{X}^d \cap \mathcal{X}^f = \emptyset$ and $\mathcal{X}^f = \mathcal{X} \setminus \mathcal{X}^d$. This dichotomous grouping of the instances is based on the values of the sensitive attributes. Besides the

sensitive attributes there are some attributes that represent the plausible reasons for preferential treatment on the basis of sensitive attributes. We refer to these attributes as explanatory attributes and denote them by E.

To illustrate the notations, consider a university where women have been denied admission in comparison to men. Here gender is a sensitive attribute (S), males belong to the favored group ($\mathcal{X}^f$), females are the deprived group ($\mathcal{X}^d$), and the acceptance or rejection decision of the selection committee defines the class label (C). Every applicant (X) who has ever applied for admission is taken as an instance of database ($\mathcal{D}$). Part of the discriminatory behavior towards women can be explained by attributes like program preference that are correlated with both the sensitive attribute and the decision. Thus, program preference is an *explanatory attribute* ($e \in E$).While selection of explanatory attributes is often debatable, we assume that they are nominated by the domain experts externally. We restrict this work to nominal explanatory attributes only.

The task is to learn a classifier $\mathcal{F}: \mathcal{X} \rightarrow \{C^+, C^-\}$ from the given discriminatory data $\mathcal{D}$ that does not make discriminatory decisions w.r.t. sensitive attribute(s) while predicting future instances. As the convention for this problem setting, the performance of the discrimination-aware classification methods is determined by reporting their accuracy and discrimination. Ideally, accuracy should suffer the least as discrimination is reduced to zero.

3.2. Measuring Discrimination

Several measures of discrimination have been proposed in the discrimination-aware classification research. In this work, we distinguish between two types of discrimination: overall and illegal discrimination. We use the definitions of [16, 6, 17, 32] for overall discrimination. Overall discrimination quantifies the difference in treatment (i.e., labelings) between deprived and favored groups on the basis of sensitive attributes only, ignoring all other explanations for the differential treatment.

Definition 1. ***(Overall Discrimination,*** D_{all}***):*** *Given a labeled dataset* $\mathcal{D} = \{X_i, C_i\}_{i=1}^N$*, sensitive attributes* S *and their respective domains describing*

instances in deprived and favored groups ($\mathcal{X}^d$ and $\mathcal{X}^f$), the discrimination in dataset $\mathcal{D}$ w.r.t. sensitive attributes S, denoted by $D_{all}(\mathcal{D}, S)$, is defined as:

$$D_{all}(\mathcal{D}, S) := \frac{|\{X \in \mathcal{X}^f, C = C^+\}|}{|\{X \in \mathcal{X}^f\}|} - \frac{|\{X \in \mathcal{X}^d, C = C^+\}|}{|\{X \in \mathcal{X}^d\}|}.$$

In probabilities, this is equivalent to $p_{\mathcal{D}}(C^+|\mathcal{X}^f) - p_{\mathcal{D}}(C^+|\mathcal{X}^d)$.

When clear from the context, we will omit the subscript and parameters in the notation, and more often, refer to this measure as overall discrimination.

Overall discrimination disregards other plausible reasons for the differential treatment between the two groups. As such, this measure is appropriate when discrimination w.r.t. sensitive attribute alone needs to be controlled (e.g., when stipulated by law).

In other scenarios, part of the differential treatment between deprived and favored groups can be explained by other attributes. For instance, low acceptance rate of female applicants to a university can be explained by their preference for more competitive disciplines (e.g., medicine). In such a scenario, discrimination that cannot be explained is called illegal discrimination. It quantifies preferential treatment on the basis of sensitive attributes without any plausible reason. We use the definition of [32] to measure illegal discrimination.

Definition 2. ***(Illegal Discrimination, $D_{illegal}$):*** *Given a discriminatory labeled dataset $\mathcal{D}$, sensitive attributes S distinguishing between instances in deprived and favored groups ($\mathcal{X}^d$ and $\mathcal{X}^f$), and explanatory attributes E. Let $dom(E) = \{1, \ldots, k\}$ be the domain of E. The explainable discrimination $D_{expl}(\mathcal{D}, S, E)$ in dataset $\mathcal{D}$ w.r.t. the sensitive attributes S and the explanatory attributes E is calculated as follows:*

$$D_{expl}(\mathcal{D}, S, E) := \sum_{i=1}^{k} \left(p(E_i|\mathcal{X}^f) - p(E_i|\mathcal{X}^d)\right) p^\star(C^+|E_i)$$

where

$$p^\star(C^+|E_i) := \frac{P(C^+|E_i, \mathcal{X}^f) + p(C^+|E_i, \mathcal{X}^d)}{2}.$$

Then, the illegal discrimination $D_{illegal}(\mathcal{D}, S, E)$ in dataset $\mathcal{D}$ w.r.t. the sensitive attributes S and the explanatory attributes E is given by:

$$D_{illegal}(\mathcal{D}, S, E) \ := \ D_{all}(\mathcal{D}, S) - D_{expl}(\mathcal{D}, S, E)$$

Here, $D_{all}(\cdot)$ is the overall discrimination in $\mathcal{D}$ as defined in Definition 1. [32].

When clear from the context, we will omit the subscript and the parameters in the notation, and more often, refer to this measure as illegal discrimination.

The above measures calculate the discrimination in any given labeled dataset. We can use the same discrimination measures to calculate the discrimination of a classifier by assuming the given dataset to be a test dataset labeled by the classifier.

4. Methodology for Discrimination Control

In this section, we present a methodology for social discrimination control that exploits the reject option in classification. The reject option in classification discards a predicted label when it is found to be highly uncertain or ambiguous. This rejection provides an opportunity for relabeling the instance in a manner that reduces discrimination while maintaining prediction accuracy over the biased dataset. We present three reject option based solutions for discrimination control: Probabilistic Rejection (PR), Ensemble Rejection (ER), and Situatinal Rejection (SR). We start by defining our discrimination model underlying the methodology.

4.1. Discrimination Model: Reject Option in Classification

Recently, a discrimination model has been presented that describes the process leading to biased labeling of instances during classification [32]. According to this model, a decision maker obtains a preliminary score m quantifying the worthiness of an individual X without relying upon the sensitive attributes describing X. Thus, this score is evaluated objectively and on merit. Then, the discrimination bias $b \geq 0$ is introduced by looking at the sensitive attributes

and their values for the individual. A uniform bias is either added (positive bias) or subtracted (negative bias) from the merit-based score m, to yield the overall score $m^* = m \pm b$. In general, the bias can vary for different individuals, however, in this study we assume a uniform bias b is added/subtracted to favor/discriminate the unprotected/protected group instances. In the social sciences, this bias is referred to as an unconscious bias [14]. The final decision of individual X is made by using score m^*.

This discriminatory decision making process impacts the decision of instances that are close to the decision boundary according to their score m. It is quite intuitive that the addition or subtraction of the bias b will not affect the decision of instances with very high or low merit-based scores m.

In our setting, we already have a discriminatory dataset $\mathcal{D}$ that captures information about the decision making process. We know key attributes of the classification problem including the sensitive attributes S, the explanatory attributes E, and the class label C. However, we do not have a clear distinction between objective or merit-based and biased contributions in the labeling process. As is required by law, the sensitive attributes cannot be used in learning and prediction. Nonetheless, because of correlation between sensitive and explanatory attributes the classifier learns the bias through the explanatory attributes. This phenomenon has been demonstrated in previous works [33].

Given the above observations, we propose the following discrimination model. Let $\mathcal{F}$ be a classifier (or a classifier ensemble) learned over the discriminatory dataset $\mathcal{D}$ without considering the sensitive attributes S, and let $0 \leq \mathcal{F}(X, C^+) \leq 1$ be the score (e.g., posterior probability or confidence) for label C^+ of instance X produced by $\mathcal{F}$ and $\mathcal{F}(X, C^-) = 1 - \mathcal{F}(X, C^+)$. Then, instance $X \in \mathcal{X}^d$ with label C^- is likely to be discriminated when $\mathcal{F}(X, C^+) \geq 0.5 - \eta$ where $0 < \eta \leq 0.5$ is a parameter that specifies the bias in the dataset. Similarly, instance $X \in \mathcal{X}^f$ with label C^+ is likely to be favored when $\mathcal{F}(X, C^+) \leq 0.5 + \eta$. Otherwise, instance X is neither discriminated nor favored according to this model.

The classifier's score $\mathcal{F}(X, C^+)$ and the parameter η correspond roughly to

m^* and b, respectively, in the basic discrimination model outlined earlier. The value of η controls the region on both sides of the classifier's decision boundary within which classification scores are considered ambiguous; instances whose scores lie in this region are not assigned a label by the classifier (i.e., their labels are rejected) and are considered likely to be the result of discriminatory practices captured in the dataset.

The parameter η can be estimated automatically when a non-discriminatory dataset is available. Alternatively, a domain expert can analyze potentially discriminated/favored instances close to the decision boundary to fix an appropriate value for η.

Definition 3. ***(Discrimination and Favoritism Potential):*** *The Discrimination Potential of an instance $X \in \mathcal{X}^d$ with label C^- in a discriminatory dataset $\mathcal{D}$ is defined as*

$$DP(X \in \mathcal{X}^d) = \mathcal{F}(X, C^+) - (0.5 - \eta) \geq 0$$

Similarly, the Favoritism Potential of an instance $X \in \mathcal{X}^f$ with label C^+ in a discriminatory dataset $\mathcal{D}$ is defined as

$$FP(X \in \mathcal{X}^f) = (0.5 + \eta) - \mathcal{F}(X, C^+) \geq 0$$

Here, $\mathcal{F}(X, C^+)$ is the score for label C^+ for instance X produced by classifier $\mathcal{F}$ learned over the discriminatory dataset $\mathcal{D}$.

$DP(\cdot)$ and $FP(\cdot)$ range from 0 to 0.5 with higher values signifying greater potential of being discriminated or favored in the dataset. The expressions for computing DP and FP can return a negative value which implies that no discrimination or favoritism exists.

This discrimination model can be used for both discrimination discovery and discrimination prevention. The Discrimination and Favoritism Potentials described above allow easy identification and ranking of instances that have potentially biased decisions in a dataset. In the following sections, we present our discrimination control solutions based on our discrimination model.

4.2. Probabilistic Rejection (PR)

Our first reject option based solution for discrimination control, called Probabilistic Rejection(PR), utilizes posterior probabilities produced by one or more probabilistic classifiers to identify instances with high label uncertainty. These instances are then labeled in a manner that neutralizes the effect of discrimination. Based on the discrimination model introduced in the previous section, PR embodies strong theoretical concepts to provide excellent control over the accuracy-discrimination trade-off for future classifications.

Before proceeding further, it is worth re-emphasizing that effective discrimination control in our setting (only discriminatory dataset available) is possible only when group membership of individuals is known. Knowledge of this information is also necessary for litigation processing and for affirmative action.

4.2.1. Labeling Strategy

Traditionally, a learned classifier assigns an instance to the class with the highest posterior probability. PR deviates from this traditional decision rule and gives the idea of a critical region in which instances belonging to deprived and favored groups are labeled with desirable and undesirable labels, respectively. We first present PR for single and multiple classifiers and then relate PR with decision theory for interpretation and control.

Consider a single classifier, and let $p(C^+|X)$ be the posterior probability for instance X produced by this classifier. When $p(C^+|X)$ is close to 1 or 0 then the label for instance X is specified with a high degree of certainty. On the other hand, when $p(C^+|X)$ is close to 0.5 then the label for instance X is more uncertain. Probabilistic rejection is adopted for all instances for which $\max[p(C^+|X), 1 - p(C^+|X)] \leq \theta$ where $(0.5 < \theta < 1)$. These instances, which lie within the *critical region*, are not assigned labels (or are labeled as 'reject'). The labels for instances in the critical region (rejected instances) are considered to be ambiguous and influenced by biases. Note that $\eta = \theta - 0.5$ relates the parameter θ with the parameter η introduced in the discrimination model.

To reduce discrimination, these rejected instances are labeled as follows; if

the instance is from the deprived group ($\mathcal{X}^d$) then label it as C^+ otherwise label it as C^-.

The instances outside the critical region are classified according to the standard decision rule, i.e., if $p(C^+|X) > p(C^-|X)$ then C^+ will be assigned to instance X; otherwise, C^- will be assigned to instance X.

Probabilistic rejection is not restricted to work with a single classifier; it can also be used for an ensemble of probabilistic classifiers. In our problem setting of discrimination-aware classification, a classifier ensemble can be thought of as a pool of experts with varying characteristics and biases – their combined output is expected to be more reliable w.r.t. both accuracy and discrimination.

Let $\mathcal{F}_k$ $(k = 1, \ldots, K)$ denote the kth classifier in an ensemble of $K > 1$ classifiers, and $p(C, \mathcal{F}_k|X)$ be the posterior probability of classification of instance X produced by classifier $\mathcal{F}_k$. The posterior probability of classification of the ensemble $p(C|X)$ is given by

$$p(C|X) = \sum_{k=1}^{K} p(C|X, \mathcal{F}_k)p(\mathcal{F}_k) \tag{1}$$

The prior probability of a classifier, $p(\mathcal{F}_k)$, can be taken to be proportional to the accuracy of that classifier on the data. Or, if such information is considered uninformative, the prior probability distribution can be taken to be uniform, in which case, the posterior probability of the ensemble is simply the average of the posterior probabilities of each classifier in the ensemble.

Given the posterior probability of an ensemble $p(C|X)$, PR proceeds in the manner as discussed for a single classifier above. This labeling strategy will ensure that only higher risk instances are rejected and thus its impact on accuracy of the classifier is a minimum. This aspect is discussed further in the next subsection.

4.2.2. Decision Theoretic Interpretation

In this section, we develop a decision theoretic understanding of PR. The expected loss of a single classifier or an ensemble of classifiers ($\mathcal{F}$) that produces posterior probabilities $p(C^+|X)$ and $p(C^-|X) = 1 - p(C^+|X)$ for instance X is

given by

$$\mathcal{E}[L] = \sum_{\{X \in \mathcal{X} | \mathcal{F}(X) = C^+\}} L_{-,+} p(C^-|X) p(X) + \sum_{\{X \in \mathcal{X} | \mathcal{F}(X) = C^-\}} L_{+,-} p(C^+|X) p(X). \quad (2)$$

Here, $L_{+,-}$ quantifies the loss incurred in classifying a positive instance as negative. These quantities are typically given in a loss matrix, with rows representing actual labels and columns giving predicted labels (Table 1). There is no loss when the predicted and actual labels match; hence, $L_{+,+} = L_{-,-} = 0$ while $L_{+,-}, L_{-,+} > 0$.

The best label for each instance X, that ensures the minimum expected loss of classification (Equation 2), is given by the $j \in \{+, -\}$ that minimizes [5]:

$$L_{+,j} p(C^+|X) + L_{-,j}(1 - p(C^+|X)). \quad (3)$$

When all classification errors incur a constant loss (e.g., $L_{+,-} = L_{-,+}$, then the above decision rule assigns each instance X to the label whose posterior probability is the largest. This is the standard decision rule that ensures the lowest loss in the accuracy of classification.

Table 1: Loss matrix

Actual, Predicted→	C^+	C^-	C^r
C^+	$L_{+,+}$	$L_{+,-}$	$L_{+,r}$
C^-	$L_{-,+}$	$L_{-,-}$	$L_{-,r}$

The reject option in classification is invoked when $max[p(C^+|X), 1-p(C^+|X)] < \theta$. From Equation 2, it is clear that even when all rejected instances (say R instances) are misclassified the increase in expected loss is a minimum as compared to any other set of R misclassified instances from a given dataset. This is because the rejected instances have a low maximum posterior probability. The labeling strategy of Probabilistic Rejection (PR), however, only relabels deprived group instances with negative labels and favored group instances with

positive labels. This strategy reduces discrimination by decreasing the dependence of the sensitive attributes on the class attribute without impacting the dependence of other attributes on the class attributes. Thus, PR reduces illegal discrimination first while maintaining the explainable discrimination.

In PR, the trade-off between accuracy and discrimination is controlled by θ; in general the larger the value of θ the greater the reduction in classifier discrimination, as more deprived and favored group instances are likely to be labeled with C^+ and C^-, respectively. For any given value of θ, the expected reduction in accuracy is the minimum possible as pointed out in the preceding paragraph. To achieve a specified discrimination level, the value of θ can be determined by using a validation dataset.

Typically in classification, a uniform cost or loss is associated with all errors, irrespective of them being false positives or false negatives. That is, $L_{+,-} = L_{-,+}$ (see Table 1), and conveniently this loss can be taken to be 1 unit. The reject option can be invoked by considering a third prediction label (C^r for reject) and taking $L_{+,r} = L_{-,r} = 1 - \theta$. Thus, the loss for rejecting an instance depends upon the value of θ – the larger its value is, the smaller the loss for rejection.

The PR labeling strategy can be interpreted via loss matrices. Consider a separate 2×2 (no C^r label) loss matrix for deprived and favored group instances (Table 2). The discrimination reducing and accuracy preserving classification is achieved when $L^d_{+,-} = L^f_{-,+} = \theta/(1-\theta)$, with the other values remaining unchanged from the usual loss matrix (Table 1).

Table 2: Loss matrices for probabilistic rejection (PR). The left matrix is for deprived instances and the right is for favored instances.

	Deprived Insts		Favored Insts	
Actual↓, Predicted→	C^+	C^-	C^+	C^-
C^+	0	$\frac{\theta}{1-\theta}$	0	1
C^-	1	0	$\frac{\theta}{1-\theta}$	0

Thus, PR can be interpreted as a cost-based prediction method in which the cost or loss of misclassifying a deprived group instance as negative is $\theta/(1-\theta)$ times that of misclassifying it as positive. A similar statement can be made for favored group instances. For example, when $\theta = 0.6$ then a 50% higher loss is associated with one type of error as compared to the other.

4.3. Ensemble Rejection (ER)

Our second reject option based solution for discrimination-aware classification, called Ensemble Rejection (ER), relabels instances on which an ensemble of classifiers disagrees significantly. Unlike PR, ER is not restricted to probabilistic classifiers only; an ensemble comprising of any type of classifier can be used in this solution. As pointed out earlier, classifier ensembles often produce robust classifications by taking advantage of the diversity of member classifiers. Furthermore, a classifier ensemble mimics practical decision making where a panel of experts converge on an outcome (e.g., acceptance or rejection) for an individual. For discrimination prevention and control, ER provides additional flexibility in the choice of a classification system.

4.3.1. Labeling Strategy

Typically, a classifier ensemble labels a new instance with the majority class label (majority-vote rule). Ensemble Rejection (ER) deviates from this standard rule to neutralize the effect of discrimination. Specifically, it labels instances on which member classifiers disagree significantly in a manner that reduces discrimination.

Formally, let $K \geq 2$ be the number of classifiers in an ensemble $\mathcal{F}$, and $0 \leq K^+ \leq K$ be the number of classifiers in the ensemble predicting label C^+ for an instance X. Then, the confidence of the C^+ label produced by the classifier ensemble $\mathcal{F}$ is defined as

$$conf(\mathcal{F}, X, C^+) = K^+/K.$$

Likewise, the confidence of the C^- label is given by $conf(\mathcal{F}, X, C^-) = 1 - conf(\mathcal{F}, X, C^+)$. Given these confidence values, ER labels instance X using

the following decision rule: if $\max[conf(\mathcal{F}, X, C^+), conf(\mathcal{F}, X, C^-)) \leq \theta$ then instance X) is assigned the desired label (C^+) if it belongs to the deprived group and the undesired label (C^-) if it belongs to the favored group. Otherwise (i.e., when $\max[conf(\mathcal{F}, X, C^+), conf(\mathcal{F}, X, C^-)) > \theta$), the standard majority-vote label is assigned to instance X.

As in PR the parameter θ, which varies from 0.5 to 1, controls the critical region in input space where the standard decision rule (majority-vote) is rejected in favor of the discrimination-aware rule to reduce discrimination. A value of $\theta = 0.5$ means that the standard majority-vote rule is utilized for all instances, while a value of $\theta = 1$ means that the majority-vote label is rejected for all instances. Thus, θ controls the trade-off between discrimination and accuracy of a specific classifier ensemble.

A special case of the ER labeling strategy is when θ is just less than one (e.g., $\theta = 0.99$). In this case, when all member classifiers predict the same label for a given instance, the agreed class label is assigned to it; otherwise, if the instance belongs to the deprived group it is assigned the C^+ label and if the instance belongs to the favored group it is given the C^- label. In other words, all instances for which the member classifiers disagree are rejected and labeled to reduce discrimination.

Based on our discrimination model, the ER labeling strategy considers that instances on which more member classifiers disagree are closer to the decision boundary and are more likely to be discriminated. We can draw a parallel between an ensemble and an admission committee: assume that some members of the committee are biased against female applicants and try to reject their applications. Hence, it is very likely that these members will only be able to affect the applicants close to the decision boundary because the highly qualified female applicants cannot be rejected due to their overall high score. If we consider member classifiers of an ensemble as admission committee members, then having more classifiers in the ensemble or increasing the acceptance confidence may neutralize the discriminatory effect of ensemble due to the fair classifiers. Thus, using classifier ensembles is a natural fit to the solution of

discrimination-aware classification problem.

4.3.2. Controlling Discrimination

There are two approaches towards controlling discrimination with ER. The first approach assumes a fixed classifier ensemble. In this approach, the trade-off between discrimination and accuracy is controlled by varying the value of θ. This approach and the corresponding discrimination-accuracy behavior is similar to that for PR.

The second approach assumes that an instance is rejected for discrimination-aware labeling whenever a given classifier ensemble disagrees on its label. In this approach, the trade-off between accuracy and discrimination is controlled by varying the composition of the ensemble. The question now is: which members should we choose and how does this impact discrimination? The accuracy-discrimination performance of a given ensemble with ER depends upon the disagreement among the member classifiers, which is defined as:

Definition 4. ***(Disagreement of a Classifier Ensemble):*** *Given a classifier ensemble $\{\mathcal{F}_k\}_{k=1}^{K}$ $(K > 1)$ trained on discriminatory dataset $\mathcal{D} = \{X_i, C_i\}_{i=1}^{N}$, the disagreement of the ensemble w.r.t. dataset $\mathcal{D}$, denoted as $disagr_{\mathcal{D}}$, is defined as:*

$$disagr_{\mathcal{D}} = \frac{|\{X_i | \exists j, k \ \mathcal{F}_j(X_i) \neq \mathcal{F}_k(X_i)\}|}{|\{X_i\}|}$$

When clear from the context, we will drop the subscript or simply use disagreement while referring to this measure.

Equivalently, $disagr_{\mathcal{D}} = d/N$, where d is the number of instances on which the ensemble disagrees. If a is the number of instances on which the ensemble agrees, then $a + d = N$. However, it is worth noticing that not all instances in a are correctly classified; the ensemble can agree on an incorrect label for an instance.

In general, the higher the disagreement of an ensemble on a given dataset, the lower will be the discrimination produced by this ensemble with ER on new

instances since the ensemble will disagree on more instances and all such instances belonging to the deprived group are labeled with C^+ and the rest are labeled with C^-. Disagreement, as defined above, can be considered to be a measure of ensemble diversity as well. Ensemble diversity has been shown to be positively correlated with ensemble accuracy determined via majority vote [21]. Another measure of ensemble diversity is average pairwise correlation between member classifiers. In [30], error bounds have been developed for classifier ensemble under reject option as a function of correlation. Therefore, a key thumb rule to remember while selecting member classifiers of an ensemble for ER is: the more diverse the member classifiers are, the higher will be the disagreement (or lower will be the correlation) among them, and the greater will be the reduction in discrimination. This means that we can control the discrimination of an ensemble with ER by changing the diversity of member classifiers. To select an ensemble with ER having a specific discrimination level, a validation dataset can be used.

4.4. Situational Rejection (SR)

Our third solution for discrimination control, called Situational Rejection (SR), combines PR or ER with a legally-grounded procedure of *situation testing*. SR includes an additional check, based on a local model of classification, for instances that are rejected and relabeled in PR or ER. As such, SR is more careful in relabeling and hence less 'brute force' in its labeling strategy. Furthermore, SR provides additional insights into the prevalence of discrimination and its control in future predictions.

4.4.1. Labeling Strategy

Situational rejection's labeling strategy for discrimination control deviates from that for PR and ER with the addition of situation testing. Situation testing or situational judgement test is a systematic procedure employed in the legal domain for determining the response of a decision maker towards an applicant's suitability for a benefit or service under different settings. In this procedure,

a hypothetical situation is assumed where a pair of applicants with similar qualifications (e.g., education, experience) but from different sensitive groups (e.g., race) apply for certain benefits (e.g., job) simultaneously. The different outcomes of such a controlled experiment can assist victims of discrimination to establish the evidence against the discriminatory practices w.r.t. certain sensitive characteristics [4, 26, 23]. Specifically, if it is found that the victim was denied the benefits while his pair was awarded the benefits then this provides evidence for the discriminatory practice.

We model situation testing via a k-nearest neighbor (k-NN) classifier [9]. This local model of classification is applied to each instance that is rejected by a probabilistic classifier or a classifier ensemble learned on the discriminatory data (i.e., the instances in the critical region produced in PR and ER). A rejected instance is compared with its neighbors and is labeled w.r.t. the majority class of its neighbors from the opposite group of sensitive attribute. For instance, a rejected female will be labeled according to majority class of the k-nearest male neighbors of this rejected female. The intuition of this method is to relabel only those rejected instances that have been treated differently as compared to their peers rather than relabeling all the rejected instances.

Since SR changes the labels of selected deprived and favored group instances in the critical region it is less 'forceful' in reducing discrimination. As such, in general, to achieve the same level of discrimination a larger critical region may be required. It is also worth noting that SR can be applied to all instances and not just to those in the critical region.

5. Experimental Evaluation

In this section, we discuss the evaluation of our methodology for discrimination control on four real-world datasets. We compare the performance of our solutions with previously proposed discrimination-aware classification methods. Since our solutions are not restricted to any specific classifier, we consider several standard classifiers for discrimination-aware classification (identifying label

of each classifier is given in parenthesis): naive Bayes (**NBS**), logistic regression (**Logistic**), k-nearest neighbor (**IBK**), and decision tree (**J48**). The first and second classifiers are generative and discriminative probabilistic classifiers, respectively, while the third is an instance-based classifier with well-defined probabilistic interpretation. We also show results with decision trees, which is an information theoretic classifier, since they have been used popularly in previous discrimination-aware classification research. Besides the above classifiers, we tried many other classifiers as well, including support vector machines (**SVM**), but do not report all results for ease of understanding.

In summary, we present and discuss the results of the following experiments for preventing overall and illegal discrimination:

1. PR: Probabilistic Rejections using single and multiple probabilistic classifiers, identified as **PR (classifier)** and **PR (1st classifier+2nd classifier+...)**, respectively.
2. ER: Ensemble Rejection with two or more classifiers, identified as **ER (1st classifier+2nd classifier+...)**.
3. SR: Situational Rejection using single and multiple probabilistic classifiers, identified as **SR (classifier)** and **SR (1st classifier+2nd classifier+...)**, respectively.
4. Comparison of our solutions' results with those of current state-of-the-art discrimination-aware classification methods, identified as **Prev Methods**.
5. Performance of our solutions (PR, ER, and SR) for illegal discrimination prevention.
6. Evaluation of PR w.r.t. different and multiple sensitive attributes.
7. Evaluation of PR on test dataset with less discrimination.

Datasets: We conduct our experiments on four real-world datasets: *Adult* [2], *Communities and Crime* [2], and *Dutch Census of 1971 and 2001* [10] datasets. Table 3 gives the important characteristics of these datasets such as number of instances, number of instances belonging to deprived group ($\mathcal{X}^d$), number of attributes in the dataset, class attribute defining the desirable and un-

desirable labels, sensitive attribute (SA), and overall discrimination (calculated using Equation 1). For experiments on less discriminatory test sets (reported in Figure 6), we change some settings in the *Dutch Census* datasets as follows: use the attribute *economic status* as class attribute rather than *occupation* as class attribute of the *Dutch Census of 2001* dataset and by removing some attributes like *current economic activity* and *occupation* from these experiments to make both datasets (Dutch 1971 and 2001) consistent w.r.t. codings. The discrimination in the *Dutch Census of 2001* dataset w.r.t. *economic status* as class attribute is 28.23%.

Table 3: Key characteristics of datasets.

Dataset	**Inst.**	$\lvert\mathcal{X}^d\rvert$	**Attr.**	**Class**	**SA**	**disc%**
Adult	16 281	5 421	14	Income	sex	19.45
Communities	1 994	1 024	122	violent criminal	race	43.14
Dutch 71	99 772	51 658	9	economic status	sex	58.66
Dutch 01	15 150	7 603	12	occupation	sex	29.85

All results reported in the paper (excluding those reported in Figure 6) are obtained using *10-fold cross-validation* and each point in the figures represents the result of an independent experiment.

5.1. Overall Discrimination Control

In this section, we show that our proposed solutions prevent effectively overall discrimination in future predictions. We also show that our proposed solutions outperform the current state-of-the-art methods over three real-world datasets (the Dutch 71 dataset is only used in Section 5.4).

5.1.1. Results of PR and SR

Figure 1 shows the results of our experiments with PR and SR (PR combined situation testing) on three datasets (labeled (a), (b), (c)). The x- and y-axis of these plots represent classifiers' discrimination and accuracy respectively, and each point is for a specific value of θ which is varied from 0.5 to a maximum

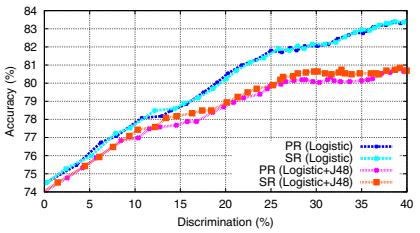
value (usually around 0.9). It is observed that as the value of θ is increased, the discrimination reduces to zero. Furthermore, the reduction in discrimination with the increase in θ is generally smooth and consistent across datasets and classifier(s). Thus, the discrimination level of PR and SR can be controlled easily by varying the value of θ. The generally small decrease in accuracy for specific values of θ makes PR and SR robust solutions appropriate for practical discrimination-aware classification.

We know that the performance of classifiers varies over different datasets; the best performing classifier over one dataset can give poor performance on another one. Figure 1 demonstrates this fact and shows that PR and SR can be used with a selected single classifier or classifier ensemble to ensure the best performances. For instance, both PR and SR give better performance with single classifiers over the Communities and Crime dataset (Figure 1 (a)). However, PR with an ensemble of logistic regression and J48 outperforms the other tested methods over the Adult dataset (Figure 1 (c)). This fact shows that the flexibility in choice of classifier(s) is really important to achieve the best results and it makes our solutions widely applicable to different domains and datasets. We can simply use the best performing classifier (single or an ensemble of multiple classifiers) on any given dataset. In general, it is seen that the classifier(s) that produces the highest accuracy at $\theta = 0.5$ for a given dataset also gives low discrimination scores by maintaining the high accuracy, making the choice of classifier(s) easier for decision makers.

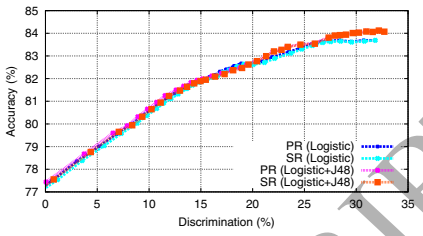
We observe in Figure 1 that both PR and SR give comparable performance. However, SR has the advantage that it can be used to establish an evidence of discriminatory practices in the court of law. This advantage of SR makes it a better choice for practitioners.

5.1.2. Results of ER

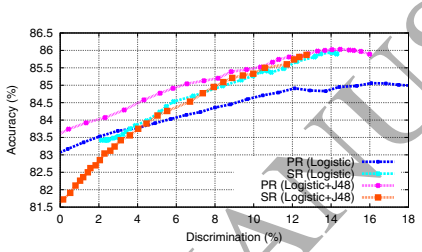
Figure 2 shows the results of our experiments with ER over three real world datasets ((a), (b), (c)). In these plots, member classifiers of different ensembles are listed on the lower x-axis, ensemble disagreement is given on the upper x-



(a) Communities and Crime



(b) Dutch Census of 2001

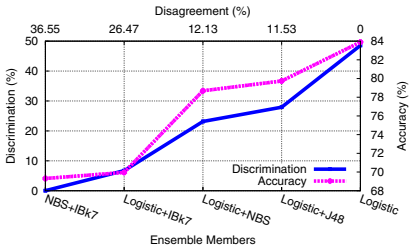


(c) Adult

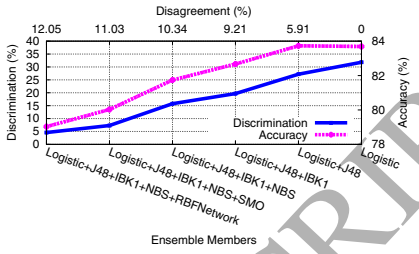
Figure 1: Discrimination-accuracy trade-off of PR and SR on three datasets. For each dataset, θ is increased from 0.5 (top right points representing standard decision boundaries) to a maximum value around 0.9 (bottom left points) which reduces the discrimination to 0%.

axis, ER discrimination is shown on left y-axis, and ER accuracy is given on right y-axis. These results demonstrate that discrimination can be controlled by varying the disagreement of the ensemble. For a given dataset, the higher disagreement the ensemble has, the lower is its discrimination with ER. The disagreement of an ensemble, which also measures the diversity of its member classifiers, can be increased by adding more classifiers. Alternatively, the disagreement can be increased by including diverse classifiers in an ensemble. For example, Figure 2 (a) shows that it is not always necessary to add more classifiers to reduce discrimination to 0%; just selecting an ensemble with high diversity (e.g., an ensemble comprising of naive Bayes (NBS) and nearest neighbor classifier with $k = 7$ neighbors (IBK7) in this case) is enough to ensure

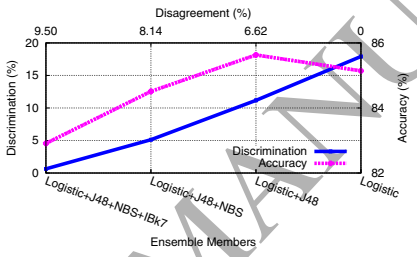
discrimination-free classification.



(a) Communities and Crime



(b) Dutch Census of 2001



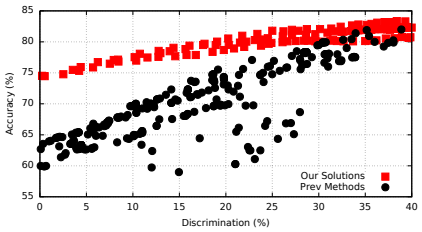
(c) Adult

Figure 2: Discrimination-accuracy trade-off of ER (disagreement based) on three datasets. For each dataset, several classifier ensembles are shown with their accuracy and discrimination.

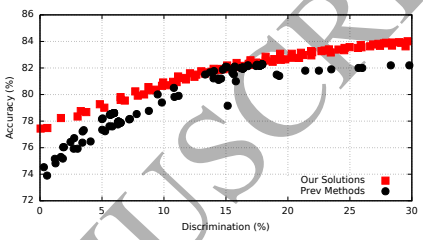
Accuracy and discrimination generally decreases with increase in disagreement. Nonetheless, accuracy remains robust since it is based on agreement of member classifiers of an ensemble. ER has an advantage that it can be used in collaboration with non-probabilistic classifiers; however, its execution time can be higher than that for PR since multiple classifiers need to be learned and applied. Similarly, SR provides a better solution for legal purposes but its execution time is the highest due to the neighborhood search step. The execution times of sample PR, ER, and SR solutions on all datasets are given in Table 4. In practice, however, execution time is not a critical deciding factor as real-world predictions do not involve stringent time constraints.

Table 4: Average execution time of PR, ER, and SR (in seconds)

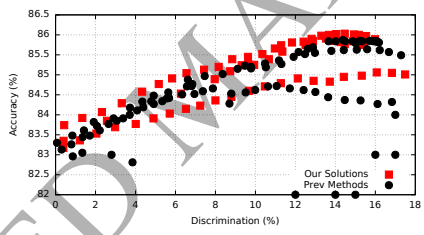
Method↓, Dataset→	Crime	Dutch	Adult
PR (Logistic)	0.58	7.86	14.23
ER (Logistic + J48)	0.76	9.33	18.54
SR (Logistic)	3.2	78	54.55



(a) Communities and Crime



(b) Dutch Census of 2001



(c) Adult

Figure 3: Comparison of our solutions with the existing state-of-the-art methods [16, 6, 17, 13, 22] on three datasets.

5.1.3. Comparison with Previous Methods

We compare the performance of our solutions (PR, ER, and SR) with that of previous methods of discrimination-aware classification. Figure 3 provides a detailed comparison of results on three real-world datasets. It is clear from the figure that our solutions outperform the previously proposed discrimination-aware classification methods of [16, 6, 17, 13, 22] w.r.t. accuracy-discrimination trade-off. For each dataset, the accuracy-discrimination curve of our methods

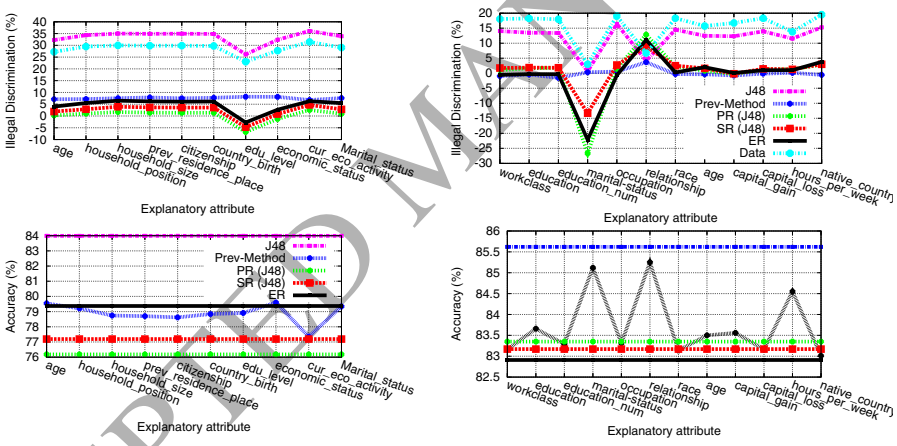
lies above all previously reported results, confirming the performance superiority of our solutions. More importantly, our solutions significantly outperform previous methods on the left side of the plots where discrimination is low but accuracy is high. To further discuss the less discriminatory results, we report highest accuracies of our proposed and previous solutions when discrimination is kept only 5%. For *communities and crime* dataset, our solutions find the highest value of accuracy (77%), while the highest accuracy of previous methods is 67% only (Figure 3(a)). A similar trend is observed for *Dutch Census of 2001* dataset, where the highest reported accuracy of our solutions is 79.2% and of previous solutions is 78.1 % (Figure 3(b)). However, the minimum difference in highest reported accuracies is discovered for the *Adult* dataset, i.e., the previous methods return 84.5% and our solutions return 84.8% (Figure 3(c)). With the increase in discrimination, the difference in the highest accuracies of our solutions and other state-of-the-arts keep decreasing, which is not justified as eventually discrimination is not prevented. These results, coupled with ease-of-use and flexible control, of our solutions make them a major step forward in practical discrimination-aware classification.

5.2. Illegal Discrimination Prevention

In this section, we empirically show that our solutions not only prevent overall discrimination but also ensure illegal discrimination prevention w.r.t. given explanatory attributes. For this purpose we present results of our experiments on two real world datasets: *Adult* and *Dutch Census*. The Communities and Crime dataset is not very appropriate for these experiments because of its small size and all numerical attributes. Although we discretize the numerical attributes in Adult and Dutch Census datasets as well but discretization of numerical attributes in Communities and Crime dataset produces very small data bins that can generate misleading results for overall and illegal discrimination.

The selection of reasonable explanatory attributes is an important step for illegal discrimination calculation and prevention. In the Adult dataset a number of attributes are very weak candidates for being explanatory attributes and

thus cannot be presented as an explanation for the low income of females. For instance, we know from biology that race and gender are independent. Thus, race cannot explain the discrimination w.r.t. gender; any such discrimination is either illegal or due to some other attributes. Similarly, the relationship attribute with values *wife* and *husband* clearly captures the gender information (i.e., is a proxy for gender) and thus cannot be used as an explanation for the low income of females. On the other hand, the attributes age and working hours per week can be considered reasonable for explaining different incomes of males and females. Therefore, it is appropriate to treat them as explanatory attributes. For Dutch Census dataset, attributes education level, age and economic activity are good candidates for explanatory attribute.



(a) Dutch Census of 2001 (b) Adult

Figure 4: Performance comparison of our solutions (PR, ER and SR) with the state-of-the-art methods of illegal discrimination prevention.

Selection of explanatory attributes is often difficult and may lead to controversies. Our solutions assume that the explanatory attributes are externally nominated (e.g., by domain experts) and in our experiments we present results by considering each attribute in the dataset as explanatory attribute.

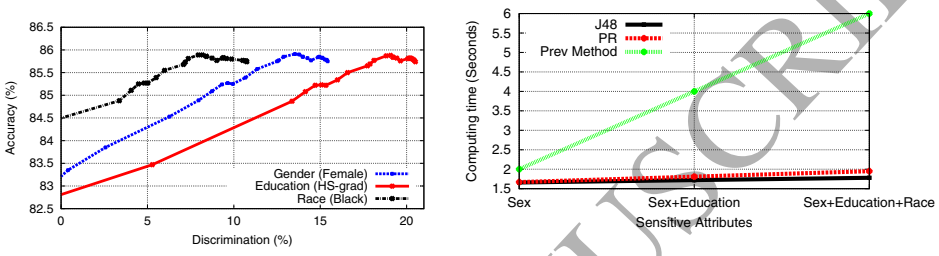
Figure 4 shows the performance of our proposed solutions w.r.t. illegal discrimination. In the plots, the x-axis shows different explanatory attributes and the y-axis shows the resultant illegal discrimination (plots on the top) and accuracy (plot in the bottom). Plots on the top of Figure 4 present the comparison of illegal discrimination in the actual data (**Data**), in the predictions of a discrimination ignorant classifier, e.g., decision tree in this figure (**J48**), and results of previously proposed methods of [32] (**Prev-Method**) with the illegal discrimination in the predictions of our proposed solutions (PR, ER, SR). We observe that our solutions reduce the illegal discrimination to almost 0% for all reasonable explanatory attributes. In general, our reject option based solutions remove the illegal discrimination with similar magnitude for all explanatory attributes as shown in Figure 4. The strange performance observed for the relationship and marital status attributes in the Adult dataset is due to the fact that these attributes are almost duplicates of the sensitive attribute (gender) and thus are not reasonable explanatory attributes, respectively.

The top two plots of Figure 4 also compare the performance of our proposed solutions with the best performing results of [32] where one specialized and independent classifier was learnt for each explanatory attribute separately. It is also very important to mention that our solutions do not require this laborious work of learning a different model for each explanatory attribute. We just learn one model to remove the illegal discrimination w.r.t. all explanatory attributes. We observe that our solutions give comparable performance with the specialized models of [32]. Our solutions are capable of reducing the discrimination to any desired level by changing the value of parameter θ. We observe even the best performing results of previous methods are not able to reduce the illegal discrimination to 0% in the Dutch Census dataset while our solutions reduce the discrimination very close to 0%.

The bottom plots of Figure 4 also give the accuracy comparison of our proposed solutions with the best performing and specialized methods of [32]. We observe that our proposed solutions give a comparable accuracy to the previous methods over the Adult dataset. However, in the Dutch Census dataset, PR

and SR are a little less accurate as they reduce the illegal discrimination to 0% as compared to the 10% range of specialized methods of [32].

5.3. Multiple Sensitive Attributes



(a) Discrimination prevention w.r.t. multiple sensitive attributes

(b) Computing time comparison with previous method [32]

Figure 5: PR's flexibility to handle discrimination w.r.t. multiple sensitive attributes without training of classification model again.

A key shortcoming of previous methods is the difficulty of handling multiple sensitive attributes which typically requires processing the data or classifier again. On the other hand, our solutions make standard classifier(s) discrimination-aware w.r.t. sensitive attribute(s) at run-time. Thus, our solutions are easy to apply to multiple sensitive attributes or different definitions of deprived groups. We demonstrate this in Figure 5(a), which shows the accuracy-discrimination trade-off of PR w.r.t. three sensitive attributes (gender, education, race) on *Adult* dataset. We observe that discrimination decreases towards zero for all sensitive attributes without repeating the learning procedure by simply increasing the value of θ from 0.5. This flexibility of PR makes it a superior discrimination-aware method as it requires very little computing resources to handle the multiple sensitive attributes as compared to other state-of-the-art methods. Figure 5(b) demonstrates this fact by comparing the computing time of PR with a standard decision tree (J48) and a previously proposed discrimination-aware

method, i.e., Massaging [16] (Prev Method) on the Adult dataset. We can observe that PR's computing time to handle discrimination w.r.t. multiple sensitive attributes is comparable to the computing time of a standard decision tree. However, the computing time of previous method becomes k times that of a single sensitive attribute when k new sensitive attributes are added, as the method has to re-run the learning process for each sensitive attribute separately. Figure 5(b) clearly points out that this drawback of previous discrimination-aware methods would become worse over large datasets.

5.4. Performance on Less Discriminatory Test Set

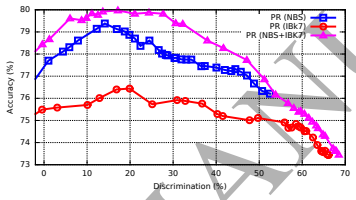


Figure 6: Performance of PR on less discriminatory test data.

Ideally, discrimination-aware classification methods trained on discriminatory data should be evaluated on discrimination-free or less discriminatory test sets. However, such evaluation scenarios are not currently available, and in state-of-the-art discrimination-aware classification research, performance is measured via accuracy-discrimination trade-off on discriminatory test sets, as reported in the previous subsections. It is expected that a discrimination-aware classifier that produces high accuracy and low discrimination on discriminatory data will perform with a higher accuracy on less discriminatory test sets. To validate this hypothesis, we construct an experiment in which PR is trained on *Dutch Census of 1971* and tested on *Dutch Census of 2001* datasets. The former dataset has a discrimination of 58.66% while the latter has a discrimination of 24.23%. As discussed while describing the datasets (Section 5), the *Dutch Census of 2001* dataset is modified to make it compatible with the *Dutch Census of 1971* dataset for this experiment, and hence, the *Dutch Census of*

2001 dataset used in previous subsections is not identical to the one used in this section.

Figure 6 shows the performance of PR using single and multiple classifiers when tested on the 2001 version after training on the 1971 version of the *Dutch Census* datasets. Unlike the results reported earlier, where both accuracy and discrimination decreases with an increase in the value of θ, here accuracy actually increases with an increase in θ from 0.5. This trend continues until discrimination is reduced to about 20%, and then accuracy starts decreasing due to the fact that the test set is not entirely discrimination free. We can expect that accuracy will continue to increase as discrimination reduces to zero if the test set is not entirely discrimination-free. This behavior of PR verifies the hypothesis and confirms its applicability to an ideal scenario where test set is less discriminatory or discrimination-free.

5.5. Summary and Discussion

Our experimental evaluations have highlighted several benefits of our proposed solutions for discrimination-aware classification. Table 5 summarizes the main advantages, relationships, and differences among the reject option based solutions. We compare our proposed solutions w.r.t. execution time, type of classifiers, and authenticity in the court of law. PR is restricted to single or multiple probabilistic classifiers, while ER and SR can use any type of classifiers. Situational Rejection (SR) is considered highly reliable for justification in the court of law, as it compares the decision of a potentially discriminated/favored instance with its neighbors to establish a case of discrimination or favoritism.

Table 5: Main features of proposed methods

Solution↓, Feature→	Non-Prob Classifier	Legal Authenticity	Run Time
PR	No	Medium	Low
ER	Yes	Medium	Medium
SR	Yes	High	High

The most significant benefit of our proposed solutions, specifically PR, is

prevention of both overall and illegal discrimination simultaneously. Actually when we increase the value of θ for PR and SR (using PR), it first removes the illegal part of discrimination and further increase of θ removes the rest of the difference in labeling between the sensitive groups to reduce the overall discrimination to zero. This benefit of our solution makes it superior to previously proposed discrimination-aware classification methods as they either reduce illegal discrimination or overall discrimination and not both. Moreover in previous illegal discrimination-aware methods, we have to learn a separate classifier for each explanatory attribute; on the other hand, our reject option based solutions prevent the discrimination w.r.t. all explanatory attributes in a single learning.

Another significant advantage of our solutions is the control over discrimination resulting from the strong correlation between θ (in PR and SR with PR) or disagreement (in ER and SR with ER) and discrimination. This kind of control is not available in the existing discrimination-aware classification methods. We have presented results for different values of θ and disagreement to establish its relationship with discrimination. In practice, if a specific discrimination level is desired, then these parameters can be fixed by using a validation dataset.

6. Conclusion

In this paper, we present three different solutions for the discrimination-aware classification problem. These easy-to-use and flexible solutions exploit the reject option in classification to identify instances to label in a manner that reduces discrimination without impacting classification accuracy significantly. The reject option in classification provides a theoretical framework for handling instances close to the decision boundary instances that are more likely to be discriminated. Our solutions employ probabilistic rejection (PR) in probabilistic classifiers, ensemble rejection in classifier ensembles (ER), and PR or ER combined with situation testing (SR). A desirable characteristic of these solutions is their interpretability, i.e., stronger justifications for the decisions as evidence against discriminatory practices in the court of law.

Our experimental evaluations on four real-world datasets confirm the benefits of our solutions and demonstrate our solutions' superior performance when compared to existing state-of-the-art methods. The results also show that our solutions prevent both overall and illegal discrimination simultaneously with minimal loss in accuracy. Stronger justifications, flexibility in practical application, ease-of-use, and overall and illegal discrimination control; these signify a major step forward in practical discrimination-aware classification.

Discrimination-aware classification is an exciting area of research with many directions for future research. Since decisions impact humans, a broader and less abstract notion of risk needs to be considered in discrimination-aware classifiers: decisions should satisfy safety requirements rather than maximizing accuracy or optimizing accuracy-discrimination trade-off [29]. Furthermore, the learned decision boundary can be quite arbitrary in low density regions thus making the use of distance from decision boundary for risk assessment more uncertain and suggesting greater human oversight in decision making [29]. We believe this direction holds much promise for future research with practical benefits. Another aspect that needs attention in discrimination-aware classification is that of causal inference where the effects of observed and unobserved explainable factors can be controlled in a systematic manner while estimating overall and illegal discrimination (e.g., [15]).

References

References

[1] ANAR (org). European Network Against Racism, 1998. via: `http://www.enar-eu.org/`.

[2] A. Asuncion and D. J. Newman. UCI machine learning repository. Online http://archive.ics.uci.edu/ml/, 2007.

[3] Canberra Attorney-General's Dept. Australian sex discimination act 1984., 1984. via: `http://www.comlaw.gov.au/Details/C2010C00056`.

[4] M. Bendick. Situation testing for employment discrimination in the united states of america. *Horizons stratégiques*, (3):17–39, 2007.

[5] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[6] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[7] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *Proc. of IEEE 13th International Conference on Data Mining*, pages 71–80, 2013.

[8] B. Custers, T. Calders, T. Zarsky, and B. Schermer. In *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 341–357. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-30486-6.

[9] A. W. David, D. Kibler, and K. M. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[10] Dutch Central Bureau for Statistics. Volkstelling, 2001.

[11] EU Legislations. European Union Legislation, 2012. via: `http://europa.eu/legislation_summaries/index_en.htm`.

[12] B. Fish, J. Kun, and Á. D Lelkes. A confidence-based approach for balancing fairness and accuracy. 2015.

[13] S. Friedler, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *CoRR*, 2014.

[14] M. Hart. Subjective decisionmaking and unconscious discrimination. *Alabama Law Review*, 56:741, 2005.

[15] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. *arXiv preprint arXiv:1605.03661*, 2016.

[16] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, pages 1–33, 2012.

[17] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *Proc. of IEEE 10th International Conference on Data Mining*, pages 869–874, 2010.

[18] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. In *Proc. of IEEE 12th International Conference on Data Mining Workshops*, pages 370–377, 2012.

[19] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *Proc. of IEEE 12th International Conference on Data Mining*, 2012.

[20] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *Proc. of IEEE 11th International Conference on Data Mining Workshops*, page 643650, 2011.

[21] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.

[22] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 502–510, 2011.

[23] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of ACM SIGKDD Conference On Knowledge Discovery And Data Mining*, pages 502–510, 2011.

[24] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.

[25] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, pages 1–57.

[26] I. Rorive. Proving discrimination cases: The role of situation testing. 2009.

[27] The US Federal Legislation. The US Federal Legislation, 2011. via: `http://www.justice.gov/crt`.

[28] UK Laws. United Kingdom Legislation, 2012. via: `http://www.legislation.gov.uk/`.

[29] K. R. Varshney. Engineering safety in machine learning. *arXiv preprint arXiv:1601.04126*, 2016.

[30] K. R Varshney, R. J Prenger, T. L Marlatt, B. Y Chen, and W. G Hanley. Practical ensemble classification error bounds for different operating points. *IEEE Transactions on Knowledge and Data Engineering*, 25(11): 2590–2601, 2013.

[31] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P Gummadi. Fairness constraints: A mechanism for fair classification. 2015.

[32] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *Proc. of IEEE 11th International Conference on Data Mining*, pages 992–1001, 2011.

[33] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. Technical report, Eindhoven University of Technology, 2011. URL `https://sites.google.com/site/conditionaldiscrimination/`.